Automatically Auditing Large Language Models via Discrete Optimization

Year: 2023 | *Citations: 206* | *Authors: Erik Jones, A. Dragan, Aditi Raghunathan*

Abstract

Auditing large language models for unexpected behaviors is critical to preempt catastrophic deployments, yet remains challenging. In this work, we cast auditing as an optimization problem, where we automatically search for input-output pairs that match a desired target behavior. For example, we might aim to find a non-toxic input that starts with"Barack Obama"that a model maps to a toxic output. This optimization problem is difficult to solve as the set of feasible points is sparse, the space is discrete, and the language models we audit are non-linear and high-dimensional. To combat these challenges, we introduce a discrete optimization algorithm, ARCA, that jointly and efficiently optimizes over inputs and outputs. Our approach automatically uncovers derogatory completions about celebrities (e.g."Barack Obama is a legalized unborn"->"child murderer"), produces French inputs that complete to English outputs, and finds inputs that generate a specific name. Our work offers a promising new tool to uncover models' failure-modes before deployment.